The Second Workshop on 3D Reconstruction Meets Semantics: Challenge Results Discussion

Radim Tylecek[1], Torsten Sattler[2], Hoang-An Le[3], Thomas Brox[4], Marc Pollefeys[2,5], Robert B. Fisher[1], Theo Gevers[4]

[1]University of Edinburgh, rtylecek@inf.ed.ac.uk
[2]Department of Computer Science, ETH Zurich, [3]University of Amsterdam
[4]University of Freiburg, [5]Microsoft, Switzerland

This discussion paper has not been peer-reviewed.

Abstract. This paper discusses a reconstruction challenge held as a part of the second 3D Reconstruction meets Semantics workshop (3DRMS). The challenge goals and datasets are introduced, including both synthetic and real data from outdoor scenes, here represented by gardens with a variety of bushes, trees, other plants and objects. Both qualitative and quantitative evaluation of the challenge participants' submissions is given in categories of geometric and semantic accuracy. Finally, comparison of submitted results with baseline methods is given, showing a modest performance increase in some of the categories.

Keywords: 3D reconstruction, semantic segmentation, challenge, dataset

1 Introduction

Over the last decades, we have seen tremendous progress in the area of 3D reconstruction, enabling us to reconstruct large scenes at a high level of detail in little time. However, the resulting 3D representations only describe the scene at a geometric level. They cannot be used directly for more advanced applications, such as a robot interacting with its environment, due to a lack of semantic information. In addition, purely geometric approaches are prone to fail in challenging environments, where appearance information alone is insufficient to reconstruct complete 3D models from multiple views, for instance, in scenes with little texture or with complex and fine-grained structures.

At the same time, deep learning has led to a huge boost in recognition performance, but most of this recognition is restricted to outputs in the image plane or, in the best case, to 3D bounding boxes, which makes it hard for a robot to act based on these outputs. Integrating learned knowledge and semantics with 3D reconstruction is a promising avenue towards a solution to both these problems. For example, the semantic 3D reconstruction techniques proposed in recent years, e.g. [9], jointly optimize the 3D structure and semantic meaning of

a scene and semantic SLAM methods add semantic annotations to the estimated 3D structure. Another recent step in this direction [5] shows that semantic and geometric relationships can be learned end-to-end from data as variational priors. Learning formulations of depth estimation, such as in [6], show the promises of integrating single-image cues into multi-view reconstruction and, in principle, allow the integration of depth estimation and recognition in a joint approach.

The goal of the 3DRMS workshop was to explore and discuss new ways for integrating techniques from 3D reconstruction with recognition and learning. In order to support work on questions related to the integration of 3D reconstruction with semantics, the workshop featured a semantic reconstruction challenge[1].

In this paper we will first present the challenge objectives and introduce datasets available for training, testing and validation of considered semantic reconstruction methods. Next, received submissions will be described, performance evaluation criteria defined and finally quantitative results will be compared and discussed.

2 Reconstruction Challenge

The challenge dataset was rendered from a drive through a semantically-rich virtual garden scene with many fine structures. Virtual models of the environment allowed us to provide exact ground truth for the 3D structure and semantics of the garden and rendered images from virtual multi-camera rig, enabling the use of both stereo and motion stereo information. The challenge participants submitted their result for benchmarking in one or more categories: the quality of the 3D reconstructions, the quality of semantic segmentation, and the quality of semantically annotated 3D models. Additionally, a dataset captured in a real garden from moving robot was available for validation.

2.1 Objectives

Given a set of images and their known camera poses, the goal of the challenge was to create a semantically annotated 3D model of the scene. To this end, it was necessary to compute depth maps from the images and then fuse them together (potentially while incorporating information from the semantics) into a single 3D model.

What we consider particularly challenging is the complex geometric structure of objects in the outdoor scenes we ask participants to reconstruct in 3D. Unlike scenes of man-made environments (indoor, urban, road-side) with certain degree of regularity of seen surfaces, a typical outdoor scene will have trees and plants with fine structures such as leaves, stems or branches, which are thin and notoriously hard to represent accurately. In real conditions those are also inherently non-rigid objects, e.g. grass moving in wind, which requires robust matching procedures to cope with small moving object parts. We hoped the participants

[1] http://trimbot2020.webhosting.rug.nl/events/3drms/challenge

would come up with representations or priors that will adapt to different objects' geometry based on their semantic class to handle such difficulties.

3 Garden Dataset

Three groups of data were provided for the challenge, see Fig. 3 for sample images.
Synthetic training sequences consist of 20k calibrated images with their camera poses, ground truth semantic annotations, and a semantically annotated 3D point cloud of 4 different virtual gardens.
Synthetic testing sequence consists of 5k calibrated images with their camera poses from 1 virtual garden.
Real-world validation sequence consists of 300 calibrated images with their camera poses from 1 real garden.
Semantic labels of objects distinguished are the following, with color code in brackets: *Grass* (light green), *Ground* (brown), *Pavement* (grey), *Hedge* (ochre), *Topiary* (cyan), *Rose* (red), *Obstacle* (blue), *Tree* (dark green), *Background* (black).

All data are available from the git repository `https://gitlab.inf.ed.ac.uk/3DRMS/Challenge2018`, where also details on the file formats can be found.

3.1 Synthetic Garden Data

We have randomly generated 5 virtual gardens (square 12m×12m) and rendered them using Blender, similar to Nature dataset [14]. The camera trajectories were generated to simulate a robot moving through the garden, moving on smooth trajectories, occasionally stopping and turning on spot, as shown in Fig. 1. At each waypoint 10 views were rendered from a virtual camera rig, which has pentagonal shape, with a stereo camera pair on each side as in Fig. 2. Fine-grained details, such as grass and leaves, were generated on the fly during rendering. Details on dataset generation can be found in [2].

3.2 Real Garden Data

The real dataset for the the 3DRMS challenge was collected in a test garden at Wageningen University Research Campus, Netherlands, which was built specifically for experimentation in robotic gardening. A validation sequence based on `test_around_garden` scenario with 124 frames from the previous year dataset was adopted for this year.

Calibrated Images. Image streams from four cameras (0,1,2,3) were provided. Fig. 2 shows these are mounted in a pairwise setup, the pair 0-1 is oriented to the front and the pair 2-3 to the right side of the robot vehicle. Resolution of the images is 752x480 (WVGA), cameras 0 and 2 are color while cameras 1 and 3 are

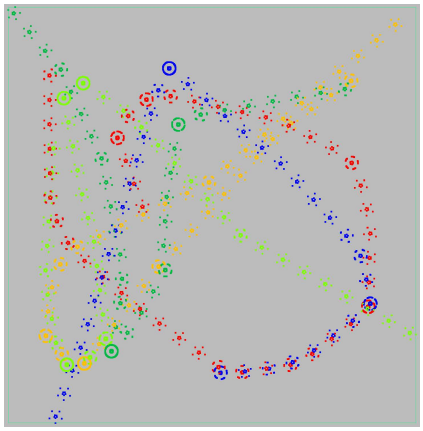


Fig. 1: Randomly generated trajectories for the test scene (unique color for each sequence)

greyscale (but sharper). All images were undistorted with the intrinsic camera parameters, calibration was performed with Kalibr toolbox [7]. The camera poses were estimated with COLMAP [17] and manually aligned to the coordinate system of the laser point cloud.

Semantic Image Annotations. Manual pixel-wise ground truth (GT) annotations (Fig. 3) produced with semantic annotation tool [20] are provided for frames from cameras 0 and 2.

Semantic Point Cloud. The geometry of the scene was acquired by *Leica ScanStation P15*, which achieves accuracy of 3 mm at 40 m. Its native output merged from 20 individual scans (Fig. 4) was sub-sampled with a spatial filter to achieve a minimal distance between two points of 10 mm, which becomes the effective accuracy of the GT. For some dynamic parts, like leaves and branches, the accuracy can be further reduced due to movement by the wind, etc.

Semantic labels were assigned to the points with multiple 3D bounding boxes drawn around individual components of the point cloud belonging to the garden objects or terrain using the `Rosemat`[2] annotation tool [20]. Ultimately the point cloud was split into segments corresponding to train and test sequences as shown in Fig. 5.

[2] Rosbag Semantic Annotation Tool for Matlab. https://github.com/rtylecek/rosemat

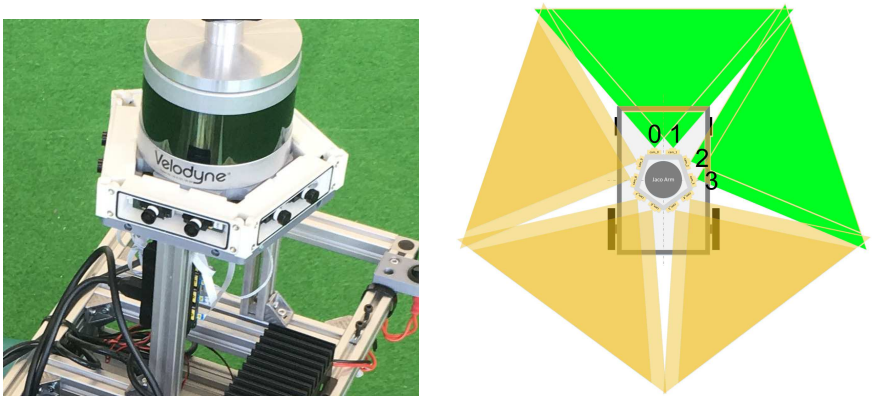


Fig. 2: Pentagonal camera rig mounted on the robot (left). First four cameras were included in the real challenge data (right, green).

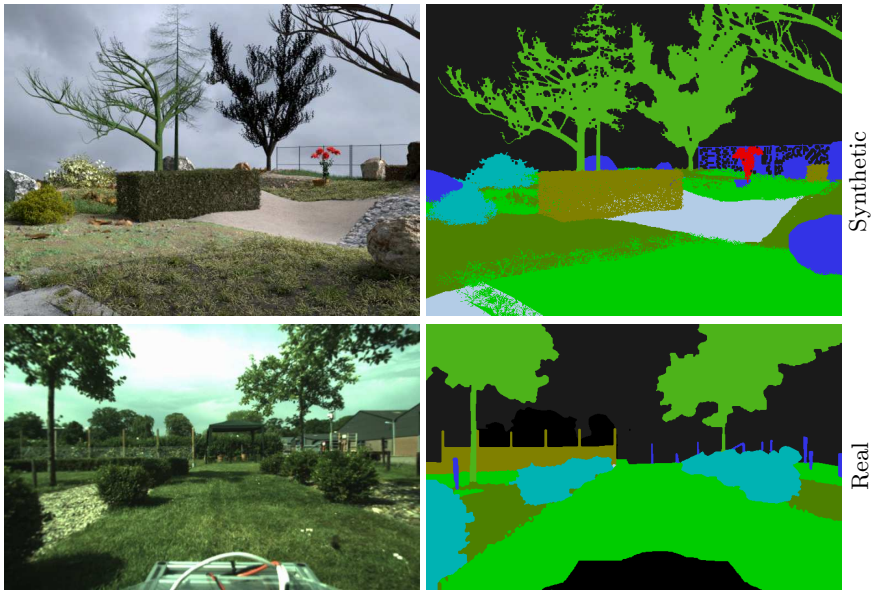


Fig. 3: Synthetic and real images of a garden from front camera mounted on a moving robot.

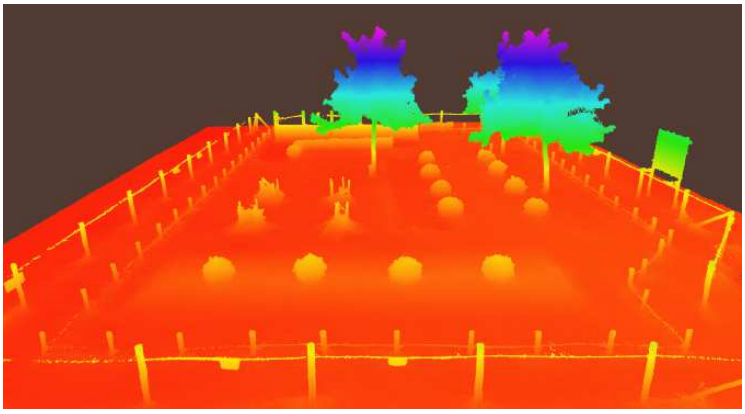
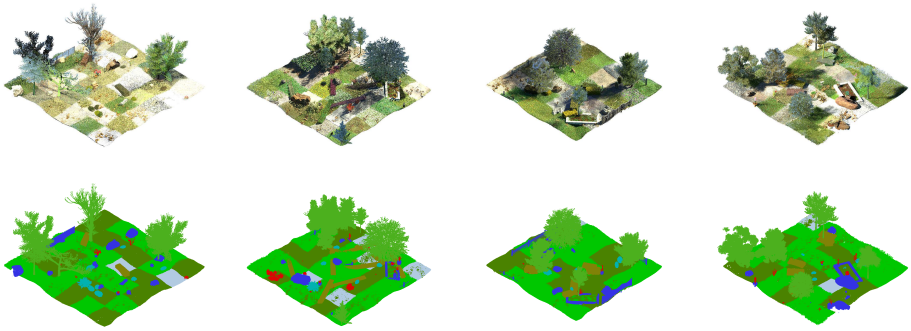
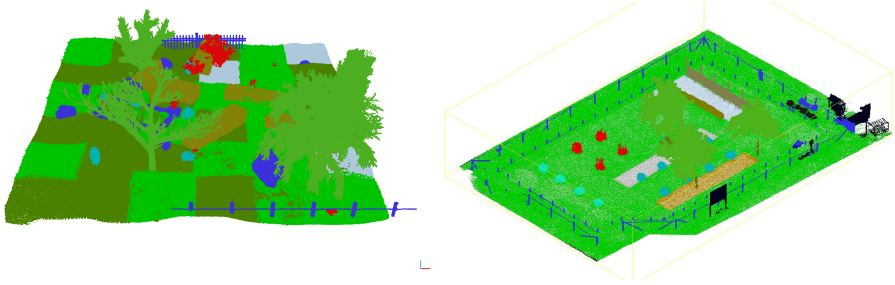


Fig. 4: Point cloud of the real garden from laser scanner (height-colored).



a) synthetic training scenes (color and semantic)



b) synthetic testing scene c) real validation scene

Fig. 5: GT semantic point cloud of virtual and real gardens with color-coded labels.

4 Submitted Results

Three submission were received fort this challenge:

DTIS [3] (ONERA, Université Paris Saclay, France): In their pipeline, initial SGM stereo results are fed to FuseNet [11], which jointly predicts a 2D semantic segmentation map and refined depth. Those are fused using TSDF in a 3D volumetric representation with colors and labels. Ultimately MC [15] extracts a surface mesh with labels assigned by voting.

HAB [10] (Video Analytics Lab, Indian Institute of Science, Bangalore, India): Their approach starts with ELAS stereo [8] producing a dense point cloud labeled with 2D semantic segmentation from DeepLabV3 [4]. The resulting point cloud is denoised with class-specific filters and similarly mesh reconstruction is using PSR [13] for flat surface classes and ball-pivoting for fine structures.

LAPSI [12] (LaPSI, UFRGS, Brazil): Only the geometric mesh was generated, in two variants: LAPSI360 using all 10 cameras and LAPSI4 using only 4 cameras. We omit the latter variant from some comparisons as it was generally performing just slightly worse than the former.

In addition to the three submitted results we have also compared to current state-of-the-art methods in both reconstruction [17] and classification [1] tasks.

COLMAP [16] (3D Reconstruction baseline) A general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. It offers a wide range of features for reconstruction of ordered and unordered image collections.

SegNet [1] (Semantic baseline) For comparison with the 2D state-of-the-art a SegNet architecture [1] is adapted for the given garden semantics.

5 Evaluation

We have evaluated the quality of the 3D meshes based on the *completeness* of the reconstruction, i.e., how much of the ground truth is covered, the *accuracy* of the reconstruction, i.e., how accurately the 3D mesh models the scene, and the *semantic accuracy* of the mesh, i.e., how close the semantics of the mesh are to the ground truth. This section describes those metrics and how we measured them.

5.1 3D Geometry Reconstruction: Accuracy & Completeness

We have followed the usual evaluation methodology described in [19]. In particular, *accuracy* is distance d (in m) such that 90% of the reconstruction is within

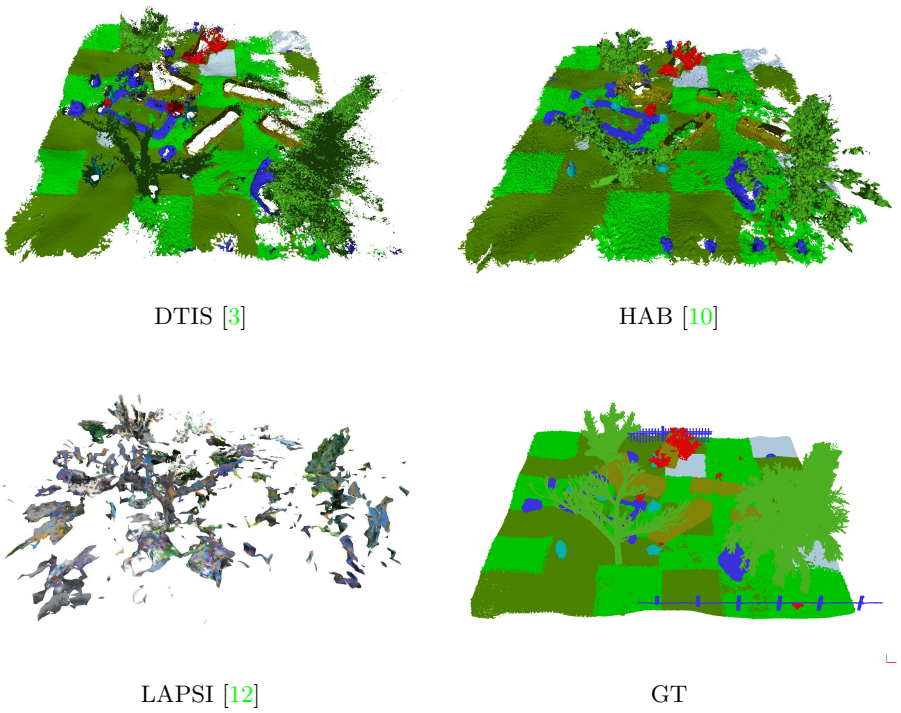


Fig. 6: Semantic and color meshes based on synthetic images submitted to the challenge with GT point cloud for comparison.

Fig. 7: Semantic and color meshes based on real images submitted to the challenge with GT point cloud for comparison.

d of the ground truth mesh and *completeness* is the percent of points in the GT point cloud that are within 5 cm of the reconstruction.

The distances between the reconstruction and GT are calculated using a point-to-mesh metric for completeness and vertex-to-point for accuracy. The faces of submitted meshes were subdivided to have a same maximum edge length. The difference between the evaluated results is shown in Fig. 8, which all use the same color scale for accuracy or completeness. Cold colors indicate well reconstructed segments while hot colors indicate hallucinated surface (accuracy) or missing parts (completeness).

The evaluation was limited to the space delimited by the bounding box of the test area plus 2 m margin. Following [18] we also plot cumulative histograms of distances in Fig. 9.

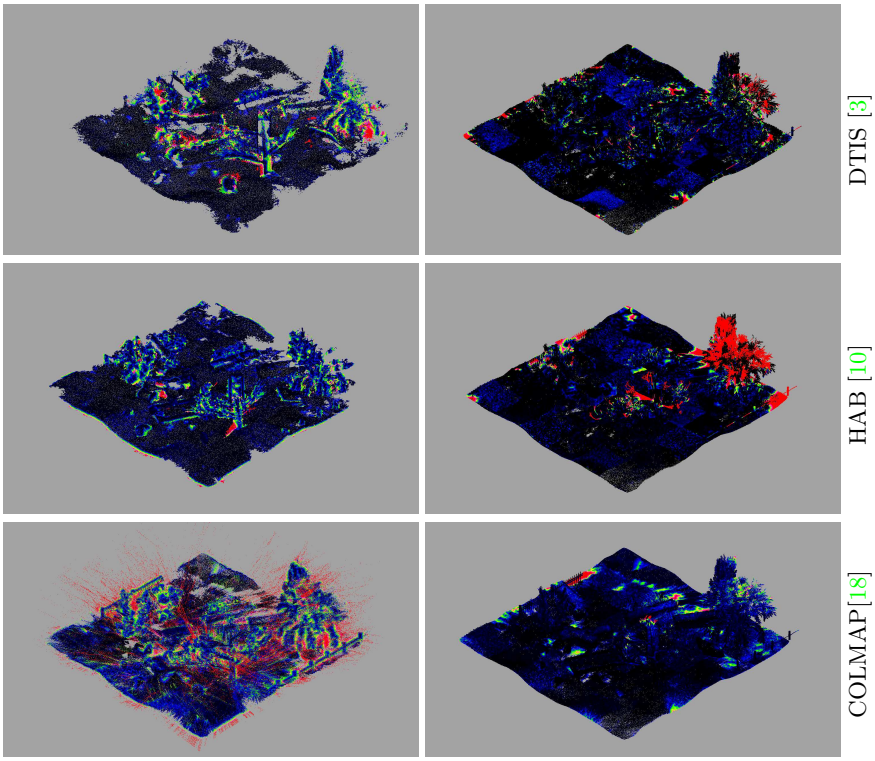


Fig. 8: Visual comparison of submitted geometry and test scene GT point cloud. Distances [0-1m]: cold colors indicate well reconstructed segments, hot colors indicate noisy surface (accuracy) or missing parts (completeness).

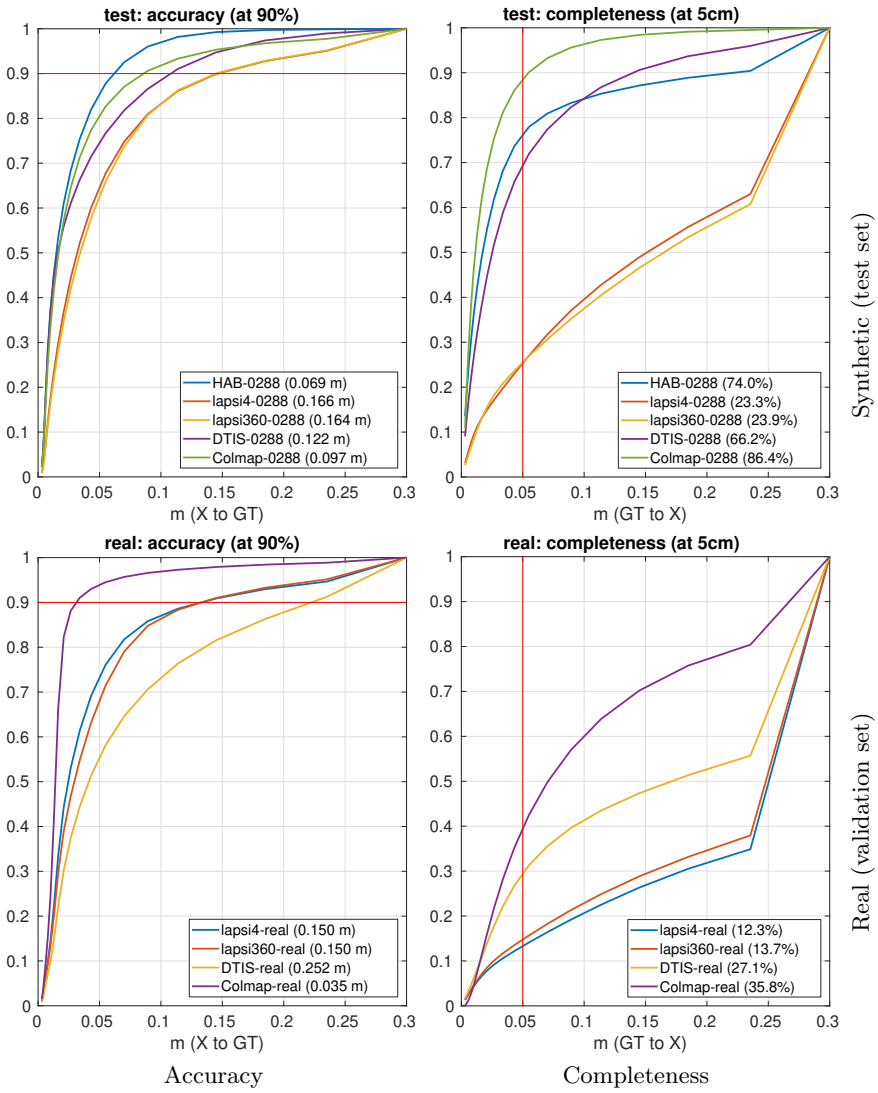


Fig. 9: Quantitative comparison of geometry with cumulative histograms of distances between GT and submissions.

5.2 Semantic Classification Accuracy

The accuracy of semantic labels assigned to vertices or faces of the 3D model (Fig. 6 and Fig. 7) was evaluated by its projection to all test images with known poses (denoted '3D' below). Some submissions also directly included image segmentation results (denoted '2D'), which were also compared.

Visual comparison of the results in a selected frame is given in Fig. 10. In the error mask the red pixels indicate incorrectly classified pixels, grey were correct and black were not evaluated. Quantitative results are presented by confusion matrices for all images in the test set in Fig. 11, where semantic accuracy is the percentage of correctly predicted pixels across all test images, and similarly in Fig. 12 for real images.

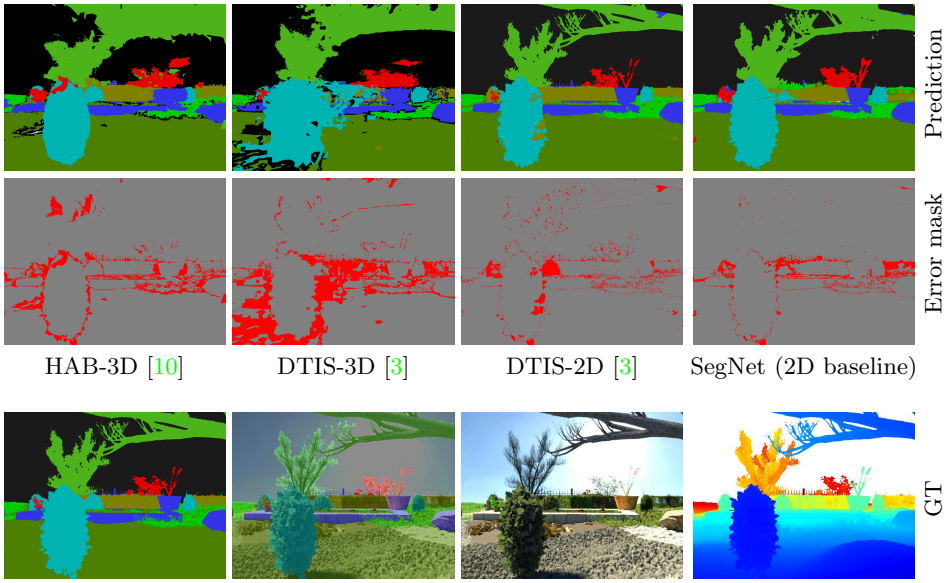
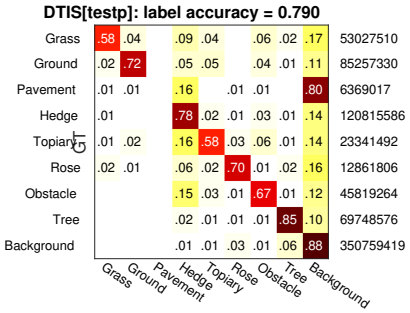
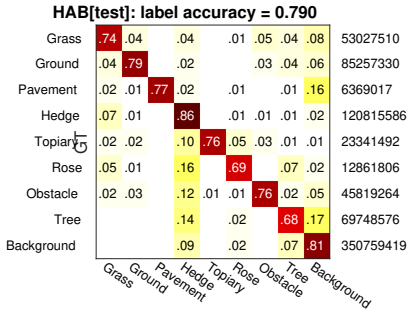


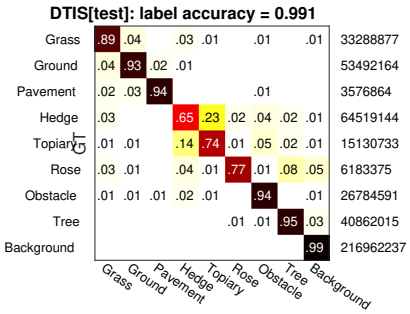
Fig. 10: Comparison of predicted semantic maps for a sample synthetic frame (above) and GT semantics with color image, overlay and depth map (below). Error mask: *red* marks incorrect pixels, *grey* correct.



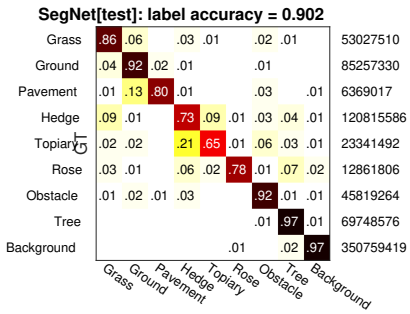
DTIS-3D 79.0%



HAB-3D: 79.0 %

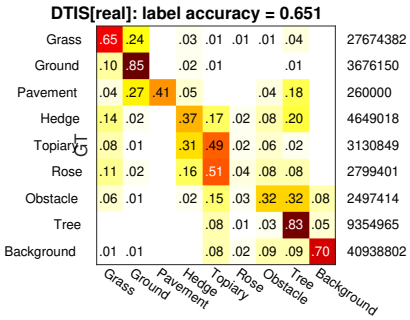


DTIS-2D: 91.1%

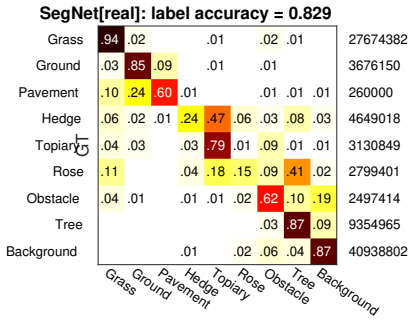


SegNet (2D baseline): 90.2%

Fig. 11: Evaluation of predicted semantic labels on test set. Confusion matrix: *dark* on diagonal indicates good match of the prediction with GT labels. Semantic accuracy: pixel-wise ratio of correct predictions over all test images.



DTIS-2D: 65.1%



SegNet (2D baseline): 82.9 %

Fig. 12: Evaluation of predicted semantic labels on real set.

5.3 Results and Discussion

The quantitative comparison in all performance categories is given in Table 1 for synthetic data and in Table 2 for real validation data.

The baseline Structure-from-Motion method COLMAP [17] was outperformed by HAB submission by 3 cm in terms of accuracy on synthetic data, but at the cost of lower completeness (Table 1). The COLMAP result could be potentially improved by filtering out outliers seen in Fig. 8, still the class-specific filters used in HAB would likely work for its advantage.

While DTIS submission was lacking good geometry, its joint depth and semantic segmentation resulted in a slight boost of 1% in 2D semantic segmentation accuracy over the SegNet baseline [1], which did not have access to depths. This however did not translate to 3D semantic accuracy, where the change of representation to less accurate mesh resulted in 12% drop in performance. Further inspection of the results shows that most object instances are correctly classified, and the 10-20% error appears near object boundaries or contours.

The real dataset proved to be more challenging Table 2, where the deep network employed by DTIS would apparently need more data for fine-tuning. This probably allowed the classic MVS baseline to prevail in both accuracy and completeness. Among the challenge participants, LAPSI was slightly better on accuracy, but their mesh was otherwise very sparse as low completeness suggests, probably resulting from overly conservative setting of the method.

In summary, best performers for synthetic data were HAB in 3D Geometry category and DTIS in the semantic category. On real data DTIS also scored better than the other submissions.

Method	3D Reconstruction		Semantic	
	Accuracy	*Completeness*	*Accuracy-2D*	*Accuracy-3D*
DTIS [3]	0.122 m	66.2 %	**91.1 %**	79.0 %
HAB [10]	**0.069 m**	**74.0 %**		79.0 %
LAPSI [12]	0.164 m	23.9 %		
Baseline	*0.097 m*	*86.4 %*	*90.2 %*	

Table 1: Comparison of submitted results on synthetic test set.

Method	3D Reconstruction		Semantic
	Accuracy	*Completeness*	*Accuracy-2D*
DTIS [3]	0.25 m	**27.1 %**	**65.1 %**
HAB [10]			
LAPSI [12]	**0.15 m**	13.7 %	
Baseline	*0.035 m*	*35.8 %*	*82.9 %*

Table 2: Comparison of submitted results on real validation set.

6 Conclusion

The workshop challenge competitors have shown that in some cases the joint semantic and 3D information reasoning can improve results. The performance gain was however rather marginal, suggesting that further optimization and design changes are needed to fully unlock the potential that such approaches offer and come up with methods giving overall balanced improvements. For this purpose, we will continue to support new authors in evaluating their methods on the garden dataset.

Acknowledgements

The workshop, reconstruction challenge and acquisition of datasets was supported by EU project TrimBot2020.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) 7, 13
2. Baslamisli, A.S., Groenestege, T.T., Das, P., Le, H.A., Karaoglu, S., Gevers, T.: Joint learning of intrinsic images and semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–302 (2018) 3
3. Carvalho, M., Ferrera, M., Boulch, A., Moras, J., Saux, B.L., Trouvé-Peloux, P.: Co-learning of geometry and semantics for online 3D mapping. In: 3DRMS Workshop Challenge, ECCV (2018) 7, 8, 9, 11, 13
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 7
5. Cherabier, I., Schönberger, J.L., Oswald, M.R., Pollefeys, M., Geiger, A.: Learning priors for semantic 3d reconstruction. In: Proc. ECCV (2018) 2
6. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proc. ICCV. pp. 2650–2658 (2015) 2
7. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: International Conference on Intelligent Robots and Systems. pp. 1280–1286 (Nov 2013) 4
8. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: ACCV 2010, pp. 25–38. Springer (2010) 7
9. Häne, C., Zach, C., Cohen, A., Pollefeys, M.: Dense semantic 3d reconstruction. Transactions on pattern analysis and machine intelligence **39**(9), 1730–1743 (2017) 1
10. Haque, S.M., Arora, S., Babu, V.: 3D semantic reconstruction using class-specific models. In: 3DRMS Workshop Challenge, ECCV (2018) 7, 8, 9, 11, 13
11. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian Conference on Computer Vision. pp. 213–228. Springer (2016) 7
12. Ilha, G., Waszak, T., Pereira, F.I., Susin, A.A.: Lapsi-360. In: 3DRMS Workshop Challenge, ECCV (2018) 7, 8, 13

13. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) **32**(3), 29 (2013) 7
14. Le, H.A., Baslamisli, A.S., Mensink, T., Gevers, T.: Three for one and one for three: Flow, segmentation, and surface normals. In: Proc. BMVC (2018) 3
15. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: ACM siggraph computer graphics. vol. 21, pp. 163–169. ACM (1987) 7
16. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision. pp. 501–518. Springer (2016) 7
17. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proc. CVPR (2016) 4, 7, 13
18. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proc. CVPR (2017) 9
19. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. CVPR. pp. 519–528. IEEE Computer Society, Washington, DC, USA (2006) 7
20. Tylecek, R., Fisher, R.B.: Consistent semantic annotation of outdoor datasets via 2d/3d label transfer. Sensors **18**(7) (2018) 4